

STUDENT EVALUATION OF TEACHING EFFECTIVENESS: METHODOLOGICAL ISSUES - PART 1

Kevin Brewer

Orsett Technical Reports, Series A, No.2

ISBN: 978-0-9540761-5-3

PUBLISHED BY
Orsett Psychological Services,
PO Box 179,
Grays,
Essex
RM16 3EW
UK

COPYRIGHT
Kevin Brewer 2002

COPYRIGHT NOTICE

All rights reserved. Apart from any use for the purposes of research or private study, or criticism or review, this publication may not be reproduced, stored or transmitted in any form or by any means, without prior permission in writing of the publishers. In the case of reprographic reproduction only in accordance with the terms of the licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of licences issued by the appropriate organization outside the UK.

General Series Introduction

Orsett Technical Reports are designed to allow the exploration of specific topics in detail. Series A contains four reports on different aspects of the student evaluation of teaching effectiveness (SETE) or students' ratings of instruction (SRI). This is the rating of lecturers and teachers by their students.

REPORT No.1 ¹

This report is a literature review of the studies into SETE and SRI, mostly from the USA. The aim is to outline what students see as the "ideal lecturer". Much of the material comes from the prolific work of Kenneth Feldman.

REPORT No.2 ²

This report addresses the issue of the accuracy of students' ratings of their instructors. Is it an accurate picture of their teaching effectiveness or the personal feelings of the students? The issues of reliability, generalisability, and validity of the ratings, along with rating errors, are discussed.

REPORT No.3 ³

Report no.3 takes many of the technical issues raised in report no.2 further. In particular, the potential biases to SETE and SRI.

REPORT No.4 ⁴

This report gives details of the construction of the Birmingham Overseas Student Teaching Evaluation Questionnaire (BOSTEQ). The aim is to produce a rating instrument specifically to be used by overseas students.

The research is part of an MSc degree at the University of Aston ⁵.

¹ Brewer, K (2002a) Student evaluation of teaching effectiveness: an introduction, Orsett Technical Reports, Series A, No.1, Orsett Psychological Services: Orsett, Essex.

² Brewer, K (2002b) Student evaluation of teaching effectiveness: methodological issues - part 1, Orsett Technical Reports, Series A, No.2, Orsett Psychological Services: Orsett, Essex.

³ Brewer, K (2002c) Methodological issues with student evaluation of teaching effectiveness (SETE) - part 2, Orsett Technical Reports, Series A, No.3, Orsett Psychological Services: Orsett, Essex.

⁴ Brewer, K (2002d) Construction of Birmingham Overseas Students Teaching Evaluation Questionnaire (BOSTEQ), Orsett Technical Reports, Series A, No.4, Orsett Psychological Services: Orsett, Essex.

⁵ Brewer, K (1993) Overseas Students Evaluation of Teaching Effectiveness, Unpublished MSc thesis, University of Aston: Birmingham, UK.

CONTENTS

	Page No.
ESTABLISHING RELIABILITY, GENERALISABILITY AND VALIDITY OF STUDENT RATING OF INSTRUCTION	4
Methods used to establish reliability	4
Generalisability of student ratings	7
Establishing validity of student ratings	8
Use of MTMM	14
Multi-section courses or MTMM design for validation?	15
ARE STUDENT RATINGS SINGLE OR MULTI-DIMENSIONAL?	16
RATING SCALES	17
RATING ERRORS ON STUDENT RATINGS OF INSTRUCTION	19
Rating error	19
Bias in student ratings	19
Implicit theories	19
Semantic similarity of items	20
Other bias	21
DOES FEEDBACK CHANGE THE TEACHING PERFORMANCE?	22
REFERENCES	23

ESTABLISHING RELIABILITY, GENERALISABILITY AND VALIDITY OF STUDENT RATINGS OF INSTRUCTION

METHODS USED TO ESTABLISH RELIABILITY

1. INTERNAL CONSISTENCY.

Using for example, odd-even or split-half, and coefficient alpha (Cronbach 1951) or Kuder-Richardson formulas (Kuder and Richardson 1937).

The aim is to correlate various questions within the instrument. Studies have shown good internal consistency. For example, Remmers and Weisbrodt (1965) Purdue Rating Scale for Instructors (PRSI) shown to have correlation of between 0.67 to 0.91 using Horst method (Horst 1949).

Costin et al (1971) quote correlations ranging from .77 to .94 for randomly paired students within a class. Feldman (1977) reports an extension of this approach, where two mean scores for a particular item can be obtained by randomly dividing a class in half. The resulting correlation is corrected by the Spearman-Brown Prophecy Formula, and it produces correlations between .70s and .90s (see Guilford 1954 for more details). Most of the commonly used instruments report reliability coefficients over 0.50. Table 1 shows a selection of post 1975 studies and the reliability coefficients reported.

But "simply computing the internal consistency of an entire questionnaire would be inappropriate unless the whole instrument were intended to measure a single quality and produce a single summary score across items" (Doyle 1975 p35).

2. TEST-RETEST.

Here the rating instrument is given to the same subjects at two different times. The aim being to correlate the two scores of each subject. For example, Remmers and Brandenburg (1927) administered the PRSI 3 days after the original use with the same group, and found a correlation between 0.42 to 0.92.

But the instructor may change between administrations of the instrument, and so a small correlation will suggest that the instrument is unstable. This method is also criticised for "being a test of the student's memory instead of being a measure of reliability" (Frey 1978 p85).

Name of Study	Instrument /Sample	Method of establishing reliability	Reliability coefficient
Frey (1978)	Endeavour 26 787 students at Northwestern University	Inter-rater	0.61 skill 0.32 rapport
Marsh (1982)	SEEQ 250 000 in 4 years at University of Southern California	(1) intraclass (2) coefficient alpha	(1) 0.74-0.90 (2) 0.88-0.97
Watkins and Thomas (1991)	SEEQ/ Endeavour combined 111 Indian students	coefficient alpha	0.54-0.93 overall 0.88 (SEEQ median) 0.87 (Endeavour median)
Fernandez and Mateo (1992)	CUTEQ-R 36 589 students at Universidad Comlutense	coefficient alpha	0.97-0.98
Watkins and Akande (1992)	SEEQ/ Endeavour combined 158 undergrads in Nigeria	coefficient alpha	0.68-0.93 overall 0.92 (SEEQ median) 0.91 (endeavour median)
Watkins and Gerong (1992)	SEEQ/ Endeavour combined 77 undergrads in Philippines	coefficient alpha	0.85-0.97 overall 0.93 (SEEQ median) 0.94 (Endeavour median)
Watkins and Regmi (1992)	SEEQ/ Endeavour combined with 297 Nepalese students	coefficient alpha	0.54-0.84 overall 0.79 (SEEQ median) 0.73 (Endeavour median)

Table 1 - showing the coefficient of reliability found by selected post 1975 studies.

3. MEAN RATINGS.

It is assumed that mean ratings of instructors should be different, because the instructors display different teaching behaviour. If the means are similar or identical, the ratings are seen as biased.

Whitely, Doyle and Hopkinson (1973) used this method in a large multi-section course; finding that the mean ratings varied between instructors.

But the assumption that instructors do differ is open to question. However, enough research has shown that the distinction between "good" and "bad" lecturers can be established (eg: Marsh 1977).

Frey (1978) used a variation of this method. He chose a sample of the data representing instructors who had taught three or more classes (with 10 + students in each), which had filled in ratings. Variance estimates were calculated for differences among instructors, and differences among classes within instructors - inter rater agreement. A formula used (recommended by Ebel 1951) showed the proportion of observed variance due to differences in instructor.

4. ANOVA.

Proposed by Guilford (1954): rather than attempting to remove potential bias, it aims to identify the contribution of bias to the final rating, and adjust for it. Obviously, this has advantages because some potential biases cannot be easily separated (eg: the halo effect).

For example, Treffinger and Feldhusen (1970), using this method, found that the halo effect only accounted for 10% of the variance in students' ratings (quoted in Doyle 1975 p43).

5. INTER-RATER RELIABILITY.

This looks at the consistency of ratings among people. Reliability here is when all raters in a group give the same pattern of responses. Usually estimated by intra-class correlation coefficients, ie: the comparison of ratings within one class of one lecturer with ratings of different instructors. Because it is sensitive to the number of raters, Centra (1979) suggests intra-class correlations of .70s for 10 raters through to .90s for 20 (p27).

Feldman (1977) makes a number of points about interpreting the reliability coefficients:

i) "reliability coefficients of individual ratings indicate the degree of general or relative consistency among raters; they do not measure exact or absolute agreement" (p229);

ii) inter-rater agreement is only the degree to which independent raters give the same rating for the same lecturer;

iii) inter-rater reliability is "the degree to which

the ratings by different raters are proportional when expressed as deviations from their means" (p229);

iv) the reliability coefficients of average college student ratings may be high, but this does not mean that individual students within the classes are highly consistent in their ratings;

v) consistency in ratings among students may not be a good basis for estimating individual ratings or average ratings reliability, particularly if the aim is to compare ratings across situations. Guthrie (1927) suggests that student ratings agree at the end of the term because of greater exposure to the lecturer, or student gossip.

GENERALIZABILITY OF STUDENT RATINGS

This is the question of whether ratings of lecturers can be compared across situations. In a detailed analysis Bausell et al (1975) compared teaching behaviours in five situations:

i) Same course taught by same instructor on two separate occasions (CS-IS).

ii) Same course taught by two different instructors (CS-ID).

iii) Different courses taught by same instructor (CD-IS).

iv) Different course taught by different instructors within the same department (CD-ID).

v) Different courses from different departments taught by different instructors (CD'-ID').

It was found that generalisation was possible with all teaching behaviour in CS-IS; with some behaviour in CS-ID and CD-IS, but not CD-ID and CD'-ID', as expected. The authors conclude that student ratings do replicate as a whole across time, even if individual items are unclear.

Table 2 shows a summary of the correlations found by selected studies.

But Smith and Cranton (1992) suggest care. They talk about the "normative assumptions" of a class, which would restrict generalizability. For example, large classes in certain subjects may see "organisation" as very important, while for smaller classes in other subjects,

STUDY	SAMPLE	CS-IS	CS-ID	CD-IS	CD-ID
Bausell et al (1975)	Unknown	.69	.33	.17	.07 (same dept) .00 (diff dept)
Feldman (1978)	2182 students	.66	.16**	.46*	-
Marsh (1982b)	8277 classes	.70	.14	.52	.06 (same dept)

* quotes mean of Hogan (1973) and Seiler et al (1977)

** quotes only Hogan (1973).

Table 2 - showing the mean correlation coefficient of student ratings in different situations by selected students.

it may be the importance of "interaction" factors. The authors emphasis that student ratings are "rather specific to the instructional setting" (p762).

More recently, Marsh and Bailey (1993) have looked at the generalisability of teaching characteristics - for example, is a lecturer who is enthusiastic, but not organised, judged the same in all courses? Using over one million SEEQ forms, during a 13 year period, the authors are happy that "instructors appear to have distinct profiles of strengths and weaknesses that are highly generalisable" (p11).

Feldman (1978) takes up the issue of whether the samples of students used in ratings are from populations of comparable raters. It is not always possible to assume that the samples are random, because students self-select themselves for courses. Thus the samples are classed as coming from a population "like those observed". This makes it possible to correlate the average class rating between two classes taught the same course by the same lecturer. The correlations are between .60s and .70s (Feldman 1978 p201).

However, the correlations leave room for other factors - for example, the course context. Feldman's (1978) conclusion is that comparison of lecturer's ratings can only be of similar sized classes, similar subjects, and similar "requiredness".

ESTABLISHING VALIDITY OF STUDENT RATING

Gaski (1987) produces evidence of studies both supporting (table 4) and non-supporting (table 3) of the

validity of student evaluations.

STUDY	SAMPLE	RESULTS
Rodin and Rodin (1972)	293 students	Inverse partial correlation between objective measure of amount learned and student rating (with initial ability controlled for)
Snyder and Clair (1976)	72 students	Expected grades inversely related to evaluations; perceived obtained grades positively related
Pratt and Pratt (1976)	175 students	Very little correlation between obtained grades and student ratings; strong positive correlation between expected grades and ratings
Brown (1976)	2360 sections; 30 000 student ratings	In stepwise regression, grades represent more powerful predictor of ratings ($r=.353$) than any other hypothesized antecedent
Powell (1977)	5 sections; 35-45 students per section	Ratings of instructor falls as grading stringency increases; amount learned increases as grading stringency increases

(Based on Gaski 1987 p327)

Table 3 - showing research generally non-supportive of validity of student evaluations.

Using the most widely accepted objective measure of validity, the student achievement test, Feldman (1989b) summarises the studies, and finds a correlation with each individual characteristic of teaching (table 5). Not surprisingly, significant correlations are found for "teacher's preparation" and "clarity and understandableness", and the student achievement test.

Braskamp et al (1985) summarise the conclusions for the use of student achievement tests as an indicator of student learning.

1. Different instructors teaching same course can be compared in terms of student performance on common exam, if classes similar in ability, prior knowledge and motivation.

2. Pre and post-course test score differences can be used to obtain an index of learning.

3. A pre-established number of students in a course

STUDY	SAMPLE	RESULTS
Gessner (1978)	78 students	High correlation between student evaluation and performance
Frey (1973)	13 instructors; 354 students	Strong relationship between student rating and teaching quality (defined as difference between observed final exam and score predicted by Scholastic Aptitude Test profile)
Marsh, Fleiner & Thomas (1975)	18 sections; 720 students	Student evaluation (across sections) positively correlated with final exam
Marsh (1977)	62 instructors; reports of most/least outstanding 591 classes; 1847 students	Evaluations validated with retrospective
Marsh, Overall & Kesler (1979)	51 instructors; 83 courses	Factor analysis indicated similar student-faculty evaluations dimensions; median $r=.49$ across evaluative factors; higher SR for courses instructor rated as most effective
Marsh & Overall (1980)	31 sections; 960 students	Generally and moderately + relationship between SR and teaching effectiveness criteria, including final exam grade (36 of 60 correlations significant)
Howard & Maxwell (1980)	Two expts: i) 8551 courses from 58 schools; 200 000 students; ii) 50 students each from 19 classes	Weak + relationship between expected grades and student satisfaction; student motivation and performance explained more of variation in satisfaction
Marsh (1982c)	329 classes	General agreement between student and instructor ratings in MTMM analysis
Howard, Conway & Maxwell (1985)	43 instructors; 34 students/ classes; 30 former students/ instructors	Student and former student ratings reported superior in convergent/discriminant validation to other methods ie: self, colleagues and trained observer ratings

(Based on Gaski 1987 p327)

Table 4 - showing research generally supportive of validity of student evaluations.

Instructional Dimension	Number of Studies	Weighted simple average correlation
1. Teacher's stimulation of interest in the course and its subject matter	14	+ .35 *
2. Teacher's enthusiasm	10	+ .26 *
3. Teacher's knowledge of subject	9	+ .24
4. Teacher's intellectual expansiveness	2	No correlation given
5. Teacher's preparation	18	+ .44 *
6. Clarity and understandableness	24	+ .47 *
7. Teacher's elocutionary skills	6	+ .33 *
8. Teacher's sensitivity to, and concern with, class level and progress	11	+ .27 *
9. Clarity of course objectives/requirements	6	+ .32
10. Nature and value of course material	10	+ .17
11. Nature and usefulness of supplementary materials and teaching aids	4	- .10
12. Perceived outcome or impact of instrument	14	+ .40 *
13. Instructor's fairness	16	+ .25 *
14. Personality characteristics	6	+ .23
15. Nature, quality and frequency of feedback from teacher	13	+ .22 *
16. Teacher's encouragement of questions and discussion	18	+ .34 *
17. Intellectual challenge and encouragement of independent thought	7	+ .23
18. Teacher's concern and respect for students	11	+ .22 *
19. Teacher's availability and helpfulness	13	+ .33 *
20. Teacher motivates students to do their best work	3	+ .33
21. Teacher's encouragement of self-initiated learning	1	- .52
22. Teacher's productivity in research	no cases	

Instructional Dimension	Number of Studies	Weighted simple average correlation
23.Difficulty in course (description)	11	+.07
24.Difficulty in course (evaluation)	9	+.05
25.Classroom management	4	+.25
26.Pleasantness of classroom atmosphere	3	+.23
27.Individualization of teaching	1	+.10
28.Instructor pursued and/or met course objectives	2	+.46 *
29.Overall rating of lecturer as an item of multi-item indicator	1	+.36
30.Overall rating of teacher as an item of multi-item indicator	3	+.38 *
31.Overall rating of course as an item of multi-item indicator	no cases	

(* = significant two-tailed $p < 0.001$)

Table 5 - showing a summary of the results of studies relating specific evaluations of teaching to student achievement as found by Feldman (1989b).

who answer correctly a specified percentage of test items can be used as an indicator of student learning (p62) ⁶.

Using construct validity requires the correlation of student ratings of a lecturer with other evaluations. Braskamp et al (1985) summarise the conclusions on lecturer self-evaluation, classroom observations by outsiders, and alumni ratings.

- Lecturer self-evaluation:

1. Students and self evaluation generally good reliability in agreement on overall ratings ⁷.
2. Agreement between students and self evaluation on dimensions of student involvement, teacher support and instructional skill ⁸.

⁶ Based on Clark (1980).

⁷ Conclusions based on Blackburn and Clark (1975); Braskamp et al (1979); Doyle and Crichton (1978); Marsh, Overall and Kesler (1979a).

⁸ Conclusions based on Braskamp et al (1980); Marsh (1980).

3. Self ratings not influenced by age, sex, tenure status, teaching load, or years of teaching experience ⁹ (p71).

- Colleagues ratings of instruction:

1. An observer may affect teaching-learning process ¹⁰.

2. Not reliable - no agreement with other methods on instructional effectiveness ¹¹.

3. The relationship between observed behaviour and student learning is not very strong ¹².

4. Colleagues ratings not highly related to student ratings, if class time was well spent and instructor open to other viewpoints ¹³.

5. Agreement between colleagues and students on specific instructional practices. They agree on descriptions of activities, but not on their judgments of instructional quality ¹⁴.

6. Colleagues are more generous than students in their ratings ¹⁵ (p66).

- Alumni ratings:

1. Same students agree between course and 1 year after graduation ¹⁶.

2. Alumni of 5 years and current students show good agreement on overall teaching effectiveness ¹⁷.

3. Alumni ratings lower than current students ¹⁸ (p74).

⁹ Conclusions based on Doyle and Webber (1978).

¹⁰ Conclusions based on Fuller and Manning (1973).

¹¹ Conclusions based on Centra (1975).

¹² Conclusions based on Braskamp et al (1985).

¹³ Conclusions based on Centra (1975).

¹⁴ Conclusions based on Centra (1975).

¹⁵ Conclusions based on Braskamp et al (1985).

¹⁶ Conclusions based on Overall and Marsh (1979).

¹⁷ Conclusions based on Centra (1974).

¹⁸ Conclusions based on Overall and Marsh (1979).

USE OF MTMM

A technique being used more and more is the multi-trait multi-method matrix (MTMM). This is basically a series of correlations of expected and unexpected behaviour with test scores.

Murphy and Davidshofer (1988) summarises three points that a test will possess as established effectively by MTMM.

- "1. Scores on the test will be consistent with scores obtained using other measures of the same construct.
2. The test will yield scores that are not correlated with measures that are theoretically unrelated to the construct being measured.
3. The method of measurement employed by the test shows little evidence of bias" (p106).

In their original article, Campbell and Fiske proposed a series of rules to follow for evaluating convergent and discriminant validity.

1. The convergent validity coefficients should be statistically significant and sufficiently different from zero to warrant further examination of the validity.
2. The convergent validities should be higher than correlations between different traits assessed by different methods.
3. The convergent validities should be higher than correlations between different traits assessed by the same method.
4. The pattern of correlations between different traits should be similar for each of the different methods (quoted in Marsh and Hocevar 1983 p233).

The above rules have been criticised. Firstly, over what constitutes a satisfactory result.

Secondly, the use of correlations based on observed variables to draw conclusions about underlying factors (Kenny and Kashy 1992 p165).

The ANOVA approach (Kavanaugh et al 1971) or the factor analysis approach (Jackson 1969) have been suggested separately to overcome the weaknesses of the MTMM matrix. However, there is not universal agreement, especially over which technique of factor analysis to use. Kenny and Kashy (1992) review a number of techniques with the MTMM matrix:

- equal loading model (Alwin 1974); all traits and methods in the matrix are allowed to correlate;
- correlated uniqueness model (Kenny 1979); no method

- factors created;
- fixed method model (Bock and Bargmann 1966); reducing one of method factors.

The authors conclude that all approaches using factor analysis have problems, and thus establishing convergent and discriminant validity is difficult.

In an earlier paper, Marsh and Hocevar (1983) compared ANOVA and confirmatory factor analysis (CFA); and recommended the latter as having specific advantages for use in the MTMM matrix.

MULTI-SECTION COURSES OR MTMM DESIGN FOR VALIDATION?

Attempts have been made to establish validity by using large multi-section courses, where different groups of students are presented the same material by different instructors.

Ideally the following controls should be used:

- many sections to the course;
- random assignment of students to the sections;
- pre-test measures used;
- each section taught by separate instructors;
- the final examination graded externally;
- common textbooks among the sections (Marsh 1984 p720).
-

Validity is then assessed by correlating the student ratings in each section.

But this does not mean a perfect methodology: each section is usually small; the problem of the influence of presage variables, like initial student motivation; the lack of consistency in measure of course achievement and student ratings. In fact Marsh goes as far as to say that this design is inherently weak (1984 p721).

Abrami, d'Appollonia and Cohen (1990) take up the defence of this methodology by reanalysing 43 multi-section validity studies. They argue that the inconsistencies of past studies were due to lack of proper analysis, and which "lacked the sensitivity necessary to identify characteristics that explain a medium size effect on the relationship between ratings and achievement" (p230).

Abrami et al (1990) point out that over 40 studies have used multi-section courses for validation of student ratings. The design is high in internal validity, allows

for some control between sections, and a common examination reduces the influence of other factors. The examination score is high in external validity - it is a direct measure of effective teaching.

Marsh (1987) advocates MTMM designs, because of the greater control on threats to internal and external validity. But to show that correlations between student ratings and, for example, instructor self-ratings are adequate measures of instruction is the problem.

ARE STUDENT RATINGS SINGLE OR MULTI-DIMENSIONAL?

Doyle (1983) originally proposed that all teaching behaviour could be covered by including 3 summary questions:

i) how would you rate this instructor's overall teaching ability?

ii) how would you rate the overall effectiveness of this course?

iii) how much have you learned as a result of this course? (Doyle 1983 p36).

The issue of whether student ratings are assessing a single or multi-dimensional behaviour in teaching became a hotly debated issue, particularly with the publication of a series of articles in the Journal of Educational Psychology in 1991. Herbert Marsh is the main proponent of a multi-dimensional approach to student ratings, while Abrami disagrees.

Marsh (1984) has no doubt that student ratings "should be unequivocally multi-dimensional (eg: a teacher may be quite well organised but lack enthusiasm)" (p709). He is against a selection of items which are then summarised by an average.

If a survey contains a hodgepodge of different items and student ratings are summarised by an average of these items or an overall rating, then there is little basis for knowing what is being measured (Marsh 1983 p151).

In a recent article, Cashin and Downey (1992) reviewing the whole debate between Marsh and Abrami, point out that a major obstacle to resolving the debate is the "lack of any agreed on criterion measure of

instrumental effectiveness" (p564). Using the Instructional Development and Effectiveness Assessment (IDEA) student rating system (Hoyt 1973), the authors are forced to accept student learning, with controls for possible bias, as the criterion of effective teaching.

The results of this study have supported that single, global items - as suggested by Abrami (1985) - can account for a great deal of the variance resulting from a weighted composite of many multi-dimensional student rating items (Cashin and Downey 1992 p569).

RATING SCALES

Doyle (1975) sees ratings composed of scales with a response mode provided (eg: agree/disagree); stems that pose a question; and cues or anchors using adjectives or phrases to define the points on a scale. In a later book, Doyle (1983) extends the various ratings that could be used to include graphic, adjectival and numerical scales; Bars (Behaviourally Anchored Rating Scales); forced choice scales; variable-item ratings; and mixed formats.

Flood Page (1974) lists examples of early rating forms, and the different systems they use. The most popular is the list of desirable teacher qualities, and the students must rate their teacher on each of them. The most common used rating is a scale with a numerical score, where 1 = "poor" to 5 = "excellent".

But how many points should be on the scale? 4 or 5 is most common. Sharpness and reliability is reduced with increasing the number of points. Wherry (1952) constructed a scale with 25 points, while Doyle (1975) recommends avoiding extremes.

An alternative to traditional scales is a double-scale format. For example, Gagne and Allaire (1974) got students to rate the instructor as they are now, and how they would want them to be. The difference between the two scores is used as an index of satisfaction or dissatisfaction.

A variation of this format involves a profile of the student's instructional needs (self ratings), and a description of what the course offers relative to satisfaction of each of those needs (Doyle 1975).

Braskamp et al (1985) summarise research on the instrumentation in student evaluation.

i) Placement of items - specific items placed before global items have a minimal effect on overall ratings. Thus global ratings can be placed at either the beginning or the end of the survey (Ory 1982).

ii) Number of response alternatives - 6 point response scales yield higher item reliabilities than 5 point response scales. Thus global items should use more 5 point response scales (Masters 1974).

iii) Negative wording of items - overall ratings are not significantly affected by number of negatively worded items. Thus both positive and negative worded items can be used (Ory 1982).

iv) Labelling all scale points vs labelling only end points - labelling only end points yields slightly higher means. Thus the response format used should be consistent for all items (Frisbie and Brandenburg 1979 quoted in Braskamp et al 1985 p45).

It is also necessary to ask whether the term "satisfied" and "dissatisfied" should be used. Peterson and Wilson (1992) show how the manner in which the question is asked about satisfaction can influence the response. They asked a question about cars; either as "how satisfied" or "how dissatisfied". The first question produced a 91% response of "very" or "somewhat satisfied", and the second only 82%. "Posing a satisfaction question in a positive form appears to lead to greater reported satisfaction than posing it in a negative form" (p65).

Questions asked earlier in the questionnaire influence subsequent answers. Peterson and Wilson (1992) found that "asking a general satisfaction question prior to a specific vehicle satisfaction question slightly increases the tendency for a 'very satisfied' response to the vehicle question" (p66).

Panney (1977) compared two versions of a rating form - one biased towards a lecturer's strengths, the other towards weaknesses. The response to the global items at the end of the rating form were as expected.

McClendon and O'Brien (1988) found that question order had an effect on questions of satisfaction with life. The placing of specific questions before general questions is important: "respondents must think about specific life domains in order to answer the general questions" (p361). For example, a general well-being question will be effected by early specific questions about marriage satisfaction.

This could have consequences for rating instruments asking specific questions, and then finishing with a question about overall evaluation of the teacher.

Schuman and Presser (1981) talk in detail about question construction on attitude surveys, including open versus closed questions; "don't knows" problem; middle position on scales; tone of wording; and order effects of questions.

RATING ERRORS ON STUDENT RATINGS OF INSTRUCTION

RATING ERROR

All ratings contain an element of measurement error. Forced-choice scales are an attempt to reduce this. The rater must choose, for example, two items from a list of four equally desirable. Sharon (1970) found this type of rating did not differ across four conditions, while a usual scale did.

But these scales have been criticised as difficult for raters, among other problems (Doyle 1975).

Research has tried to identify student characteristics that could bias ratings of teachers (reviews by Feldman 1977; 1978; 1979). When correlations between characteristics and ratings are large, there is seen to be bias, and the ratings lack validity. But Marsh (1987) argues that validity is lost only when "biasing" characteristics influence the ratings, and not the instructional effectiveness criteria at the same time, and vice versa.

BIAS IN STUDENT RATINGS

Two areas of bias that have particularly concerned researchers are the effect of implicit theories ("halo effect"), and the semantic similarity of items.

Implicit Theories

This is the idea that if the raters notice certain characteristics in the teacher, then they assume the teacher must also have certain others. Whitely and Doyle (1976) feel that students' implicit theories influence their rating of teaching. Using latent partition analysis (Wiley 1967), they identified latent clusters of behaviour, when the unit of analysis was total-class, within-class or between-class ratings. But the authors

feel that the implicit theories are based on experience of teaching, and so are quite accurate to which behaviours are associated together.

Larson (1979) is critical:

we still have no way of knowing whether a particular set of behaviour ratings reflects the actual behaviour of those being rated or whether they reflect population based normative assumptions about these behaviours (p210).

More recently, Widmeyer and Loy (1988) replicated Kelley's (1950) "first impressions" experiment finding that subjects who were told that the lecturer had a warm personality rated them as a more effective teacher, than subjects who were told the lecturer had a cold personality. The "warm personality" was also seen as less unpleasant, more sociable, less irritable, less ruthless, more humorous, less formal and more humane (p119). Full details of the results in table 6.

Marsh (1987) points out that implicit theories can be ruled out by establishing a factor structure similar to that of SRI from another method. For example, the use of lecturer's self evaluation. This method suffers from little or no "halo effect", while colleagues' evaluations may suffer most.

Item: Teaching ability	"Warm" Group	"Cold" Group	Signif- icance
Knows his material - doesn't	1.44	1.65	.05
Considerate of class - self centred	1.90	2.24	.01
Intelligent - unintelligent	1.59	1.84	.05
Organised - not	1.77	1.73	NS
Expresses himself well			
- difficulty	2.09	2.10	NS
Interesting - boring	2.05	2.32	.05

7 point scale used: 1 = left hand end to 7 = right hand end.

(Based on Widmeyer and Loy 1988 p120)

Table 6 - showing mean ratings given to a stimulus person designated warm or cold.

Semantic Similarity of Items

This is slightly different to the implicit theories, in that the raters score items because they appear to be similar to other items. For example, lecturers who are "friendly towards individual students" will be assumed to

also make "students feel welcome in seeking help/advice". Cadwell and Jenkins (1985) found evidence of this process using a hypothetical instructor profile with 28 graduate students. They explain the cognitive processes involved in responding to a SRI, which by its nature must lead to this bias.

Thus, because student ratings are the product of cognitive processes that reconstruct rather than mirror instructor behaviour, these ratings, like all personality assessments, lead us to view behaviour as more organised and more consistent than it actually is (p392).

This study has been criticised at length (Marsh and Groves 1987), particularly because of the use of a "hypothetical profile".

Again, this bias can be eliminated by construct validation.

Other Bias

1. "SIMPLISTIC BIAS HYPOTHESIS".

This states that if an instructor gives high grades, demands little work, or teaches only small classes, they will receive a higher rating. Marsh (1987) quotes his own earlier research, which he believes clearly refute this, and showed it to be a "strawman" (p310). The use of the Student Educational Evaluation Questionnaire (SEEQ), and multi-dimensions to the ratings, reduces the possibility of a global item influenced by the above factors. Furthermore, the dimension of Workload/Difficulty was opposite to this "hypothesis".

2. LENIENCY ERRORS.

The tendency to rate generously for those people the rater is involved with. Centra (1975) found that colleagues' ratings (mean of 4.47 out of 5) of teaching effectiveness was one standard deviation higher than students' (mean of 3.98). Other studies have found slightly different results. Doyle (1983) feels that "some degree of leniency error can be expected in most evaluation" (p75), but it is higher for colleagues' evaluations.

DOES FEEDBACK CHANGE THE TEACHING PERFORMANCE?

It is generally felt that student feedback should improve teaching performance. Tuckman and Oliver (1968) show that high school teachers who received feedback were later rated higher than those teachers who did not receive feedback. But Miller (1971) found no significant difference in the end of term ratings between those teaching assistants who received mid-term feedback and those who did not.

Remmers (1959) makes two important points:
"1. Knowledge of student opinions and attitudes leads to improvement of the teacher's personality and educational procedures.
2. Students are more favourable to student ratings than instructors, but more instructors have noticed improvement in their teaching as a result of student ratings than the studies have"
(Quoted in Flood Page 1974 p68).

Wilson (1986) details a scheme to improve faculty teaching at the University of California, Berkeley. Improvement was found on nine characteristics of half of the lecturers due to feedback. The technique used was consultation with lecturer using student comments.

Marsh (1987) reviews the two main type of studies aimed at answering the question of the effect of feedback on teaching.

i) Short term feedback studies - generally it is felt that feedback and consultation can improve teaching. Cohen's (1981) meta-analysis of feedback studies found that instructors receiving mid-term feedback were rated higher than the control group on overall rating. L'Hommedieu et al (1990) point out the problem of the "John Henry effect", ie: teachers who know they are being rated tried to improve their teaching.

ii) Long term feedback studies - Marsh feels that there are so few studies, and many problems with such research, that it is difficult to reach a conclusion.

In a recent study, Marsh and Roche (1993) looked at the effect of feedback from students and consultation mid-term and end of term, on the evaluation at the end of the course. The ratings improved for both groups, but only ratings for the end of term group improved significantly more than the control group (no feedback). The authors conclude that "SET (student evaluation of

teaching) feedback coupled with consultation is an effective means to improve teaching effectiveness" (p 217).

Ryan et al (1980) looked at SRI from the point of morale of the faculty. Over 90% of the 193 academics felt morale had "greatly or somewhat decreased" through the use of student ratings of their teaching. Furthermore, nearly 45% reported a decrease in job satisfaction, and over 70% a decrease in their confidence in the administration. Many academics admit their behaviour has changed to some extent because of the ratings. But the authors are sceptical of the benefits, suggesting that the most frequently reported change was a reduction in coursework demands on the students. On most behaviours, "no change" was the most often response.

REFERENCES

- Abrami, P.C; d'Apollonia, S & Cohen, P.A(1990) Validity of student ratings of instruction: what we know and what we do not know, *Journal of Educational Psychology*, 82, 2, 219-231
- Alwin, D.F(1974) Approaches to the interpretation of relationships in the MTMM matrix. In Costner, H.L (ed) *Sociological Methodology 1973-4*, San Francisco: Jossey-Bass
- Bausell, R.B; Schwartz, S & Purchit, A(1975) An examination of the conditions under which various student rating parameters replicate across time, *Journal of Educational Measurement*, 12, 4, 273-280
- Blackburn, R,T & Clark, M.J(1975) An assessment of faculty performance: some correlates between administrators, colleagues, students and self-ratings, *Sociology of Education*, 48, 242-256
- Bock, R.D & Bargmann, R.E(1966) Analysis of covariance structures, *Psychometrika*, 31, 507-534
- Braskamp, L; Caulley, D.N & Costin, F(1979) Student ratings and instructor self-ratings and their relationship to student achievement, *American Educational Research Journal*, 16, 295-306
- Braskamp, L; Brandenburg, D; Kohen, E; Ory, J & Mayberry, P(1980) *Guidebook for Evaluating Teaching*, Urbana, Ill: University of Illinois
- Braskamp, L; Brandenburg, D & Ory, J(1985) *Evaluating Teaching Efficiency: A Practical Guide*, Beverley Hills, CA: Sage
- Brown, D.L(1976) Faculty ratings and student grades: a university-wide multiple regression analysis, *Journal of Educational Psychology*, 68, 5, 573-578
- Cadwell, J & Jenkins, J(1985) Effects of the semantic similarity of items on SRI, *Journal of Educational Psychology*, 77, 4, 383-393
- Cashin, W.E & Downey, R.G(1992) Using global student rating items for summative evaluation, *Journal of Educational Psychology*, 84, 4, 563-572
- Centra, J.A(1974) The relationship between student and alumni ratings of teachers, *Educational and Psychological Measurement*, 34, 321-325
- Centra, J.A(1975) Colleagues as raters of classroom instruction, *Journal of Higher Education*, 46, 327-337

- Centra, J.A(1980) Determining Faculty Performance, San Francisco: Jossey-Bass
- Cohen, P.A(1981) Student ratings of instruction and student achievement: a meta-analysis of multi-section validity studies, Review of Educational Studies, 51, 3, 281-309
- Costin, F; Greenough, W.T & Menges, R.J(1971) Student ratings of college teaching: reliability, validity and usefulness, Review of Educational Research, 41, 511-535
- Cronbach, L.J(1951) Coefficient alpha and the internal structure of tests, Psychometrika, 16, 297-334
- Doyle, K.O(1975) Student Evaluation of Instruction, Lexington, Mass: Lexington Books
- Doyle, K.O(1983) Evaluating Teaching, Lexington, Mass: Lexington Books
- Doyle, K.O & Crichton, L.A(1978) Student, peer and self-evaluation of college instruction, Journal of Educational Psychology, 70, 815-826
- Doyle, K.O & Webber, P.C(1978) Self-ratings of college instruction, American Educational Research Journal, 15, 467-476
- Feldman, K.A(1977) Consistency and variability among college students in rating their teachers and courses: a review and analysis, Research in Higher Education, 6, 223-274
- Feldman, K.A(1978) Course characteristics and college students' ratings of their teachers: what we know and what we don't know, Research in Higher Education, 9, 199-242
- Feldman, K.A(1979) The significance of circumstances for college students' ratings of their teachers and courses, Research in Higher Education, 10, 2, 149-172
- Feldman, K.A(1989b) The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multi-section validity studies, Research in Higher Education, 30, 6, 583-646
- Fernandez, J & Mateo, M.A(1992) Student evaluation of university teaching quality: analysis of a questionnaire for a sample of university students in Spain, Educational and Psychological Measurement, 52, 3, 675-686
- Flood Page, C(1974) Student Evaluation of Teaching: The American Experience, London: Society for Research in Higher Education
- Frey, P.W(1973) Student ratings of teaching: validity of several rating factors, Science, 182, 83-85
- Frey, P.W(1978) A two-dimensional analysis of student ratings of instruction, Research in Higher Education, 9, 69-91
- Frisbie, D.A & Brandenburg, D.C(1979) Equivalence of questionnaire items with varying response formats, Journal of Educational Measurement, 16, 43-48
- Fuller, F.F & Manning, B.A(1973) Self-confrontation review: a conceptualisation for video playback in teacher education, Review of Educational Research, 43, 469-528
- Gagne, F & Allaire, D(1974) Summary of Research Data on the Reliability and Validity of a Measure of Dissatisfaction Derived from Reality-Desires Discrepancies, Quebec: Institute National de la Recherche Scientifique, Universite de Quebec
- Gaski, J.F(1987) On "construct validity of measures of college teaching effectiveness", Journal of Educational Psychology, 79, 3, 326-330

- Gessner, P.K(1973) Evaluation of instruction, *Science*, 180, 566-570
- Guilford, J.P(1954) *Psychometric Methods* (2nd ed), New York: McGraw-Hill
- Guthrie, E.R(1927) Measuring student opinion of teachers, *School and Society*, 25, 175-176
- Kavanagh, M.J; MacKiney, A.C & Wolins, L(1971) Issues in managerial performance: MTMM analyses of ratings, *Psychological Bulletin*, 75, 34-49
- Kelley, H.H(1950) The warm-cold variable in first impressions of persons, *Journal of Personality*, 18, 431-439
- Kenny, D.A(1979) *Correlation and Causality*, New York: Wiley
- Kenny, D.A & Kashy, D.A(1992) Analysis of the MTMM matrix by confirmatory factor analysis, *Psychological Bulletin*, 112, 1, 165-172
- Kuder, G.F & Richardson, M.W(1937) The theory of the estimation of test reliability, *Psychometrika*, 2, 151-160
- Larson, J.R(1979) The limited utility of factor analytic techniques for the study of implicit theories in student ratings of teacher behaviour, *American Educational Research Journal*, 16, 2, 201-211
- L'Hommedieu, R; Menges, R.J & Brinko, K.T(1990) Methodological explanations for the modest effects of feedback, *Journal of Educational Psychology*, 82, 232-241
- McClendon, M.J & O'Brien, D.J(1988) Question order effects on the determinants of subjective well-being, *Public Opinion Quarterly*, 52, 351-364
- Marsh, H.W(1977) The validity of students' evaluations: classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors, *American Educational Research Journal*, 14, 4, 441-447
- Marsh, H.W(1980) The influence of student, course and instructor characteristics in evaluations of teaching, *American Educational Research Journal*, 17, 219-237
- Marsh, H.W(1982a) SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching, *British Journal of Educational Psychology*, 52, 77-95
- Marsh, H.W(1982b) Factors affecting students' evaluations of the same course taught by the same instructor, *American Educational Research Journal*, 19, 485-497
- Marsh, H.W(1982c) Validity of student evaluation of college teaching: a MTMM analysis, *Journal of Educational Psychology*, 74, 264-279
- Marsh, H.W(1984) Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility, *Journal of Educational Psychology*, 76, 5, 707-754
- Marsh, H.W(1987) Students' evaluations of university teaching: research findings, methodological issues, and directions to future research, *International Journal of Educational Research*, 11, 253-388
- Marsh, H.W & Bailey, M(1993) Multidimensional students' evaluations of teaching effectiveness: a profile analysis, *Journal of Higher Education*, 64, 1, 1-18
- Marsh, H.W & Groves, M.A(1987) Students' evaluations of teaching effectiveness and implicit theories: a critique of Cadwell and Jenkins (1985), *Journal of Educational Psychology*, 79, 4, 483-489
- Marsh, H.W & Hocevar, D(1983) Confirmatory factor analysis of MTMM

matrices, Journal of Educational Measurement, 20, 3, 231-248

Masters, J.R(1974) The relationship between number of response categories and reliability of Likert-type Questionnaires, Journal of Educational Measurement, 11, 49-53

Miller, M.T(1971) Instructor attitudes towards and their use of, student ratings of teachers, Journal of Educational Psychology, 62, 235-239

Murphy, K.R & Davidshofer, C.O(1988) Psychological Testing: Principles and Applications, Englewood Cliffs, NJ: Prentice Hall

Ory, J.C(1982) Item placement and wording effects on overall ratings, Educational and Psychological Measurement, 42, 767-775

Overall, J.U & Marsh, H.W(1979) Midterm feedback from students: its relationship to instructional improvement and students' cognitive and affective outcomes, Journal of Educational Psychology, 71, 856-865

Panney, G(1977) An Experiment to Investigate Possible Instructor-Induced Item Response Bias in the Purdue Cafeteria Instructional Evaluation System, Paper presented at the 7th Annual South East American Institute for Decision Sciences, Alabama

Peterson, R.A & Wilson, W.R(1992) Measuring customer satisfaction: fact or artefact, Journal of the Academy of Marketing Sciences, 20, 1, 61-71

Powell, R.W(1977) Grades, leniency and student evaluation of instruction, Research in Higher Education, 7, 193-205

Pratt, M & Pratt, T(1976) A study of the student-teacher grading interaction process, Improving College and University Teaching, 24, 73-80

Remmers, H.H(1959) The Appraisal of Teaching in Large Universities, Ann Arbor: University of Michigan

Remmers, H.H & Brandenberg, G.C(1927) Experimental data on the Purdue Rating Scale for Instruction, Educational Administration and Supervision, 13, 519-527

Remmers, H.H & Weisbrodt, J.A(1965) Manual of Instructions for the Purdue Rating Scale for Instructors (Rev ed), West Lafayette, Ind: University Book Store, Purdue University

Rodin, M & Rodin, B(1972) Student evaluation of teachers, Science, 177, 1164-1166

Ryan, J.J; Anderson, J.A & Birchler, A.B(1980) Student evaluation: the faculty responds, Research in Higher Education, 12, 4, 317-333

Schumann, H & Presser, S(1981) Questions and Answers in Attitude Surveys, New York: Academic Press

Seiler, L.H; Weybright, L.D & Stang, D.J(1977) How Useful are Published Evaluation Ratings Selecting Courses and Instructors, Unpublished manuscript

Sharon, A.T(1970) Eliminating bias from student rating of college instructors, Journal of Applied Psychology, 54, 278-281

Smith, R.A & Cranton, P.A(1992) Students' perceptions of teaching skills and overall effectiveness across instructional settings, Research in Higher Education, 33, 6, 747-764

Snyder, C.R & Clair, M(1976) Effects of expected and obtained grades on teacher evaluation and attribution of performance, Journal of Educational Psychology, 68, 75-82

Treffinger, D.J & Feldhusen, J.F(1970) Predicting students' ratings of instruction, Proceedings, 78th Annual Convention, American Psychological Association, 621-622

- Tuckman, B.W & Oliver, W.F(1968) Effectiveness of feedback to teachers as a function of source, *Journal of Educational Psychology*, 59, 297-301
- Watkins, D & Akande, A(1992) Student evaluations of teaching effectiveness: a Nigerian investigation, *Higher Education*, 24, 453-463
- Watkins, D & Gerong, A(1992) Evaluating undergraduate college teaching: a Filipino investigation, *Educational and Psychological Measurement*, 52, 3, 727-734
- Watkins, D & Regmi, M(1992) Student evaluations of tertiary teaching: a Nepalese investigation, *Educational Psychology*, 12, 2, 131-142
- Watkins, D & Thomas, B(1991) Assessing teaching effectiveness: an Indian perspective, *Assessment and Evaluation in Higher Education*, 16, 3, 185-198
- Wherry, R.J(1952) Control of Bias in Ratings, Department of the Army, The Adjutant General's Office, Personnel Research and Procedures Division, Personnel Research Branch, PRS Reports 914, 915, 919, 920 and 921
- Whitely, S.E & Doyle, K.O(1976) Implicit theories in student ratings, *American Educational Research Journal*, 13, 4, 241-253
- Widmeyer, W.N & Loy, J.W(1988) When you're hot: warm-cold effects in first impressions of persons and teaching effectiveness, *Journal of Educational Psychology*, 80, 1, 118-121
- Wiley, D.E(1967) Latent partition analysis, *Psychometrika*, 32, 183-193
- Wilson, R.C(1986) Improving faculty teaching, *Journal of Higher Education*, 57, 2, 196-211